View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Kent Academic Repository

Kent Academic Repository

Full text document (pdf)

Citation for published version

Everett, Richard and Nurse, Jason R. C. and Erola, Arnau (2016) The Anatomy of Online Deception: What Makes Automated Text Convincing? In: 31st Annual ACM Symposium on Applied Computing (SAC).

DOI

https://doi.org/10.1145/2851613.2851813

Link to record in KAR

http://kar.kent.ac.uk/67493/

Document Version

Author's Accepted Manuscript

Copyright & reuse

Content in the Kent Academic Repository is made available for research purposes. Unless otherwise stated all content is protected by copyright and in the absence of an open licence (eg Creative Commons), permissions for further reuse of content should be sought from the publisher, author or other copyright holder.

Versions of research

The version in the Kent Academic Repository may differ from the final published version.
Users are advised to check **http://kar.kent.ac.uk** for the status of the paper. **Users should always cite the published version of record.**

Enquiries

For any further enquiries regarding the licence status of this document, please contact:
researchsupport@kent.ac.uk

If you believe this document infringes copyright then please contact the KAR admin team with the take-down information provided at **http://kar.kent.ac.uk/contact.html**

KAR Kent Academic Repository

University of Kent

The Anatomy of Online Deception: What Makes Automated Text Convincing?

Richard M. Everett
Cyber Security Centre,
Department of Computer Science,
University of Oxford, UK
richard.everett@cs.ox.ac.uk

Jason R.C. Nurse
Cyber Security Centre,
Department of Computer Science,
University of Oxford, UK
jason.nurse@cs.ox.ac.uk

Arnau Erola
Cyber Security Centre,
Department of Computer Science,
University of Oxford, UK
arnau.erola@cs.ox.ac.uk

ABSTRACT

Technology is rapidly evolving, and with it comes increasingly sophisticated bots (i.e. software robots) which automatically produce content to inform, influence, and deceive genuine users. This is particularly a problem for social media networks where content tends to be extremely short, informally written, and full of inconsistencies. Motivated by the rise of bots on these networks, we investigate the ease with which a bot can deceive a human. In particular, we focus on deceiving a human into believing that an automatically generated sample of text was written by a human, as well as analysing which factors affect how convincing the text is. To accomplish this, we train a set of models to write text about several distinct topics, to simulate a bot's behaviour, which are then evaluated by a panel of judges. We find that: (1) typical Internet users are twice as likely to be deceived by automated content than security researchers; (2) text that disagrees with the crowd's opinion is more believably human; (3) light-hearted topics such as Entertainment are significantly easier to deceive with than factual topics such as Science; and (4) automated text on Adult content is the most deceptive regardless of a user's background.

CCS Concepts

•**Human-centered computing** → **Social networking sites;** *User studies;* •**Security and privacy** → Social network security and privacy;

Keywords

Social Media, Social Bots, Human Factors, Reddit

1. INTRODUCTION

Social media networks are being increasingly overwhelmed by software robots (known as bots) which automatically generate content to inform, influence, and deceive genuine users [5]. While not all of the bots on these sites are malicious, weather bots for example, those that are often, try to hide the fact that they are not human. As technology improves, these bots – and their underlying algorithms – will become more sophisticated, making it harder to distinguish their content from that produced by real users. This is particularly a problem for social media networks where users interact through posting short snippets of informally written text which may not be in reply to anything specific. This problem is further aggravated with social media networks now being used in marketing [15], financial trading [3], and social uprisings [11]. The result of this is that it gives the owners of the bots – be they individuals, organisations, or even governments – the ability to deceive users and influence their opinions [19], often without the users recognising it.

Motivated by the rise of bots on social media networks, and their apparent success [12], the goal of this research is to assess the ease with which a bot can deceive a human using automatically generated text (automated text). In particular, deceiving a human into believing that the text generated was written by a human, and understanding which factors affect how convincing the text is. We focus on the Topic of the text, its length, as well as how other users view and rate the content of the text (called Crowd opinion). For each of these, results are compared between typical Internet users and security researchers. We envisage that this research on these factors can be used to assist and educate Internet users in the detection of automatically generated content, reducing the chance of users being deceived. As far as we are aware, this work is one of the first attempts in research that seeks to investigate these factors in more detail.

To achieve our goal, we train several statistical models to generate text about five distinct Topics (Information, Science, Entertainment, Humour, and Adult) with three different Crowd opinions (Positive, Negative, and Neutral). Each model is then used to generate text which is combined with unique samples of real text to form a test dataset. The success of the models is then evaluated by a number of judges who each label every sample in the test dataset as either genuine (human) or automatically generated (bot).

Overall, we found that automatically generated text can deceive security researchers more than 25% of the time, and typical Internet users more than 50% of the time. Also, disagreeing with the crowd (negative Crowd opinion) makes automated text look more human, increasing the likelihood of deception by up to 45% compared to the average, and up to 78% when compared to text that agrees with the crowd. Furthermore, text on a light-hearted Topic such as Entertainment is up to 85% more likely to deceive than that on a factual Topic such as Science. Finally, automated text on

Adult content is the most deceptive for both typical Internet users and security researchers, increasing the likelihood of deception by at least 30% compared to other Topics.

The remainder of this paper is organised as follows. To begin, the related work on social bots and automated text is presented in Section 2. Next, the research and experiment setup to analyse the factors affecting what makes automated text convincing is described in Section 3. This is followed by a presentation of the results and discussion related to the hypotheses in Section 4, after which a reflection is given in Section 5. The paper is then concluded in Section 6 alongside an outline of future work.

2. RELATED WORK

Online deception is not a new concept [8], however it is evolving, and quickly. As exemplified in the Ashley Madison data leak, the online dating website that specialises in people seeking extramarital relationships, bots are starting to automatically perform online deception on a large scale [12]. While bots have existed since the early days of computers, technological and social developments in recent years have led to the rise of sophisticated bots, particularly on social media networks. Known as 'social bots', they can automatically produce content and interact with users [10], and they are becoming increasingly sophisticated.

With this rise of social bots, researchers have started to investigate which factors make users susceptible to the automated content that bots produce. Using the data from the 2011 Social Bot Challenge, Wagner *et al.* [17] modelled the susceptibility of users on Twitter to automated content about cats, finding that susceptible users tended to be more socially active. Similarly, Wald *et al.* [18] examined the attributes of users which interacted with Twitter Social Bots, finding that users with a large number of friends and high Klout Score[1] were most susceptible to automatically generated content. More recent work by Dickerson *et al.* [7] found that humans express sentiment in their text on Twitter more strongly than bots, and that humans tend to disagree more with the general Twitter population.

Academic work has shown that it is possible for a social bot to function effectively. For example, Boshmaf *et al.* [4] operated a social botnet for 8 weeks and demonstrated that social networks – such as Facebook – could be infiltrated by a bot with a success rate of up to 80%. In 2013, Zhang *et al.* [20] demonstrated the effectiveness of social botnets for spam distribution and influence manipulation through real experiments on Twitter.

There also exists several documented examples of humans mistaking automatically generated text for genuine text outside of social media. One of the more infamous recent examples is that of certain conferences and journals accepting automatically generated publications [13], an area which is meant to be under high levels of scrutiny. Similar algorithms are also starting to write news articles, from short summaries to more extensive reports, and are becoming more common on news websites [9]. In 2014, Clerwell [6] examined whether respondents could tell software-generated text from that written by a journalist. While they only used one generated sample of text, they found that 37% of respondents thought a journalist wrote it.

Unlike academic papers and news articles, content on social media networks tends to be extremely short, informally written, and full of inconsistencies. Twitter[2], for example, has a strict 140 character limit on Tweets which encourages shorthand writing, and two thirds of the comments on Reddit[3] are shorter than 140 characters even though the limit is 10,000. These features make it easier to deceive using automated text, opening up various research questions about the factors which make users susceptible to the automated content that bots produce. As far as we are aware, this is one of the first attempts in research that seeks to investigate the factors which influence a user based on Topic, Crowd opinion, and the user's background.

3. RESEARCH DESIGN

To reiterate, the aim of this paper is to investigate the ease with which a bot can deceive a human into believing that a sample of automated text was written by a human, as well as to understand which factors affect how convincing the text is. In this section, we describe the data used to accomplish this aim, along with our hypotheses and the experiment.

3.1 Dataset

The dataset used in our work consists of user comments from Reddit, a community-driven platform for submitting, commenting on, and rating content. Receiving more than 200 million unique visitors per month[4], it is considered one of the largest online communities on the Web and boasts prominent users such as Barack Obama and Bill Gates.

Two distinct advantages of using Reddit as a data source are that (1) it consists of thousands of sub-communities (known as 'subreddits') which discuss specific Topics, and (2) it allows users to vote on comments to change its score, whereby comments with higher number of votes are promoted and placed more prominently on the site than those with lower votes. Both of these attributes are included in the metadata of each comment, and allow us to investigate our hypotheses described in Section 3.2.

To access user comments on Reddit, we use the publicly available 'Complete Public Reddit Comments Corpus' [1] which contains 99.98% of all comments from October 2007 to May 2015. After removing the unnecessary fields and filtering out non-human comments, the remaining data is formatted as shown in the following example:

[text:'I want the truth to this', score:9, category:'science']

3.2 Hypotheses

The hypotheses designed to guide our research and the experimentation were as follows:

Hypothesis 1 - Topic: The Topic of a comment has an impact on its ability to convince a reader that it was written by a human. In particular, factual Topics tend be more difficult for a bot to automatically write about, and therefore easier for humans to identify correctly, while more light-hearted Topics are easier to write about and therefore more convincing. To investigate this, comments on two factual Topics (Information and Science) and two light-hearted Topics (Entertainment and Humour) are include in our experiments. A fifth Topic, for comments about Adult content,

[1]https://klout.com/corp/score

[2]https://www.twitter.com

[3]https://www.reddit.com

[4]https://www.reddit.com/about, as of 12th Sep 2015

is also included due to the prevalence – and apparent success – of adult bots on the Internet [12]. These Topics are described in more detail below:

1. Information: Comments on facts, skills, and the asking and answering of questions.
2. Science: Comments discussing scientific areas such as physics and medicine.
3. Entertainment: Comments about books, games, television and movies.
4. Humour: Comments on jokes and humorous content.
5. Adult: Comments discussing mature content.

Hypothesis 2 - Crowd Opinion The opinion that other users (i.e. the crowd) have on a comment has an impact on the comment's ability to convince a reader that it was written by a human, even when the Crowd opinion is not explicitly displayed. To investigate this, we assign each comment to one of three groups based on its score; Positive, Negative, and Neutral. These are described as follows:

1. Positive: Comments with a score above 0. A majority of the crowd agree with, approve of, or appreciate the content of the comments and expressed that through voting them positively (called 'upvoting' on Reddit).
2. Negative: Comments with a score below 0. A majority of the crowd disagree with or disapprove of the content of the comments and expressed that through voting them negatively (called 'downvoting' on Reddit).
3. Neutral: Comments with a score equal to 0. The crowd has a neutral opinion on the content of the comments and expressed that through an equal number of positive and negative votes. Also includes comments with no votes which represent no Crowd opinion due to restrictions in Reddit's metadata.

Hypothesis 3 - Length The length of a comment has an impact on its ability to convince a reader that it was written by a human. Due to the complexity of automatically producing consistent text about a Topic over several sentences, automated text will be easier to identify correctly the longer it is. Additionally, genuine text will be easier to identify the longer it is for because humans do not have this difficulty.

Hypothesis 4 - Reader's Background Most importantly, the background of the user reading the comment will have an impact on how successful they are at identifying automatically generated text from genuine text. In particular, technically literate and security-aware users will find it easier to identify automated text than typical Internet users.

3.3 Model

To investigate the ease with which automatically generated text can deceive humans, we have chosen to use a Markov chain model [14] to generate text on particular Topic / Crowd opinion pairings. Due to its simplicity, both to understand and create, any resulting findings will represent a lower bound on what is possible in the space of online bot-related deception, especially for a determined actor with resources such as a controlling government.

A Markov chain generates text using a collection of probabilities which are pre-computed from a training set (which we create in the next section). As an example, the first order Markov chain trained on the sentence 'the way the wind blows' is presented in Table 1. For this work, we use a second order Markov chain which calculates the probability for the next word using both the current and previous word. While a Markov chain's simplicity means that it does not learn any high level features about language, and can therefore produce text which does not always make sense, for our initial study this is acceptable given our feasibility context.

Table 1: Trained first order Markov chain

Current Word	Next Word (Probability)
the	way (0.5), wind (0.5)
way	the (1.0)
wind	blows (1.0)

3.4 Experiment Setup

In order to evaluate each model, and investigate the hypotheses pertaining to the factors which affect how convincing text is, we designed the experiment described here.

Using the dataset introduced in Section 3.1, we produced 15 subsets through the pairing of (i) Topic (Information, Science, Entertainment, Humour, Adult) and (ii) Crowd Opinion (Positive, Negative, Neutral), each containing 250,000 randomly sampled comments by real users on Reddit. For example, the 'Science/Positive' subset only contains comments about the Science Topic with a Positive score.

After producing the 15 subsets, a model is trained on each (one model per subset) and then used to generate 10 comments. These 10 comments are added to the test subset, as well as a further 10 randomly sampled comments written by real users from the corresponding Topic/Crowd opinion subset. This process is illustrated in Figure 1.

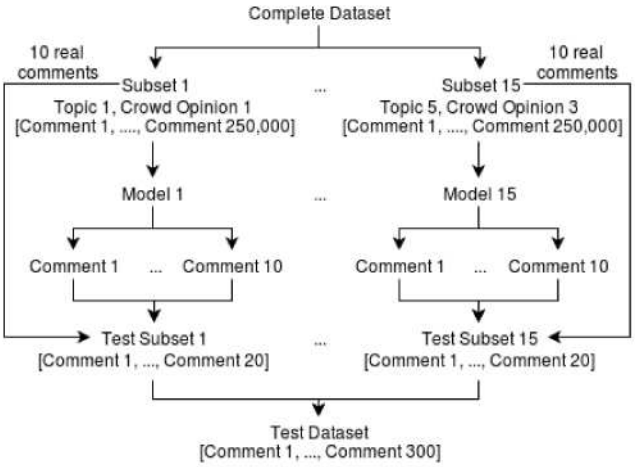

Figure 1: The process of creating the test dataset

The 15 test subsets are then combined and randomly permuted to form one complete test dataset which contains all 300 comments – 150 genuine and 150 generated.

To evaluate this test dataset, a panel of judges is used where every judge receives the entire test set with no other accompanying data such as Topic and Crowd opinion. Then, each judge evaluates the comments based solely on their text and labels each as either human or bot, depending who they believe wrote it. To fill this panel, three judges were

selected – in keeping with the average procedure of the work highlighted by Bailey *et al.* [2] – for two distinct groups:

- Group 1: Three cyber security researchers who are actively involved in security work with an intimate knowledge of the Internet and its threats.
- Group 2: Three typical Internet users who browse social media daily but are not experienced with technology or security, and therefore less aware of the threats.

4. RESULTS AND DISCUSSION

In this section, the results of our experiment are presented and discussed in the context of our hypotheses.

4.1 Metrics

Three different metrics are used to quantify the factors which affect how convincing a sample of text is. They are:

Deception Rate The percentage of automatically generated comments which are incorrectly labelled by a judge as written by a human. In the tables containing this metric, 'Factual Average' and 'Light-Hearted Average' are added which contain the average of the results from the group's judges for the Information / Science and Entertainment / Humour Topics respectively. The average of all five topics is included at the bottom of the tables.

False Rate The percentage of real comments which are incorrectly labelled by a judge as written by a bot.

Correlation (r) This metric is calculated using Pearson's correlation coefficient, a standard measure for quantifying the strength of linear association between two variables. We compare one feature of a comment with its probability of being labelled correctly by a judge, producing a value between -1 and +1 where +1 is total positive correlation, 0 is no correlation, and -1 is total negative correlation.

These metrics are calculated for each judge and then averaged to give the mean and standard deviation for each group. We compare our findings using relative difference in percentages, followed by the two compared percentages.

4.2 Topic

In Table 2, the deception rate results of security researchers (Group 1) and typical Internet users (Group 2) are presented for every Topic and their averages. These results show that automatically generated comments about light-hearted Topics, such as Entertainment and Humour, are more likely to deceive than both the average comment and comments about factual Topics like Information and Science.

Table 2: Deception rate grouped by Topic

Topic	Group 1	Group 2
Information	14.4% ± 2.0%	41.1% ± 5.1%
Science	18.9% ± 10.7%	32.2% ± 1.9%
Entertainment	33.4% ± 11.0%	58.4% ± 1.8%
Humour	28.6% ± 10.7%	59.5% ± 2.1%
Adult	40.0% ± 10.0%	67.8% ± 1.9%
Factual Average	16.7% ± 7.3%	36.7% ± 6.0%
Light-Hearted Average	31.0% ± 10.0%	58.9% ± 1.8%
Average	27.1% ± 12.6%	51.8% ± 13.8%

For Group 1, the security researchers, comments on a light-hearted Topic were more likely to deceive than those on a factual Topic by a factor of 85.6% (16.7% to 31.0%), and increased the deception rate by 14.4% from the average's 27.1% to 31.0%. Similarly, for Group 2, the typical Internet users, comments on a light-hearted Topic were more likely to deceive than those on a factual Topic by a factor of 60.5% (36.7% to 58.9%), and increased the deception rate by 13.7% from the average's 51.8% to 58.9%. This suggests that certain topics are easier to deceive with than others.

Notably, automatically generated comments on Adult content were the most difficult for both groups to label correctly. Group 1 incorrectly labelled 40.0% of the comments (a factor of 48% above the average's 27.1%), while Group 2 incorrectly labelled 67.8% (a factor of 31% above the average's 51.8%). This provides evidence that adult social bots, which are increasingly prevalent online, may be one of the most effective bots for targeting users regardless of the user's background. It also lends support to the claim that a majority of the 'women' on Ashley Madison could have been bots [12].

From these results, we can conclude that the Topic of a comment does appear to have an impact on its ability to convince users that it was written by a human, thus positively supporting *Hypothesis 1.*

In Table 3, the false rate results of security researchers (Group 1) and typical Internet users (Group 2) are presented for every Topic and their averages.

Table 3: False rate grouped by Topic

Topic	Group 1	Group 2
Information	8.9% ± 9.6%	31.1% ± 7.0%
Science	5.5% ± 6.9%	47.9% ± 7.7%
Entertainment	10.4% ± 3.6%	38.5% ± 1.8%
Humour	11.9% ± 5.5%	33.9% ± 3.1%
Adult	12.2% ± 5.1%	13.3% ± 3.4%
Factual Average	7.2% ± 7.7%	39.4% ± 11.3%
Light-Hearted Average	11.2% ± 4.2%	36.2% ± 3.4%
Average	9.8% ± 6.0%	32.9% ± 12.5%

Group 2 was more than three times as likely to believe that a real comment was written by a bot than Group 1 (32.9% vs 9.8%). This suggests that security researchers, likely due to their background knowledge, knew what to look for and were more familiar with the Internet than Group 2, providing evidence for *Hypothesis 4.* A notable exception to this is that both groups were equally capable of correctly labelling genuine Adult comments as human (12.2% vs 13.3%).

There is a clear disparity between Group 1 and Group 2 in labelling comments correctly. Group 2 was deceived nearly twice as often as Group 1 (51.8% vs 27.1%), while incorrectly labelling real comments three times as often (32.9% vs 9.8%). Both of these provide evidence to support *Hypothesis 4* that the reader's background and prior knowledge has a strong impact on their ability to correctly label comments. It also means that there exists some understanding which can be taught to typical Internet users to help them to discern automatically generated content from genuine content.

4.3 Crowd Opinion

In Table 4, the deception rate results of security researchers (Group 1) and typical Internet users (Group 2) are presented for each Crowd opinion and their average. These results show that automatically generated comments which have a negative Crowd opinion are more likely to deceive than the average comment.

Table 4: Deception rate grouped by Crowd opinion

Crowd Opinion	Group 1	Group 2
Positive	22.0% ± 8.7%	45.3% ± 2.3%
Negative	39.3% ± 15.0%	58.7% ± 1.2%
Neutral	20.0% ± 2.0%	51.3% ± 3.1%
Average	27.1% ± 12.7%	51.8% ± 6.1%

For Group 1, comments with a negative Crowd opinion were more likely to deceive than those with a positive Crowd opinion by a factor of 78.6% (22.0% to 39.3%), and increased the deception rate by 45.0% from the average's 27.1% to 39.3%. Similarly, for Group 2, comments with a negative Crowd opinion were more likely to deceive than those with a positive Crowd opinion by a factor of 29.6% (45.3% to 58.7%), and increased the deception rate by 14.4% from the average's 51.3% to 58.7%. An explanation for this result is that humans tend to disagree more with the general population than bots, as found by Dickerson *et al.* [7], and therefore when a bot does disagree it is more believably human.

These results provide support for *Hypothesis 2* that the Crowd opinion on an automatically generated comment has an impact on how convincing it is. In particular, comments with a negative Crowd opinion have a higher deception rate, while those with a positive Crowd opinion actually have a lower deception rate.

In Table 5, the false rate results of security researchers (Group 1) and typical Internet users (Group 2) are presented for each Crowd opinion and their average.

Table 5: False rate grouped by Crowd opinion

Crowd Opinion	Group 1	Group 2
Positive	10.0% ± 6.9%	26.7% ± 2.3%
Negative	8.7% ± 3.1%	28.7% ± 5.0%
Neutral	10.7% ± 5.0%	42.7% ± 5.0%
Average	9.8% ± 4.6%	32.7% ± 8.4%

Similar to the findings for Topic, Group 2 was more than three times as likely to misclassify a genuine comment than Group 1 (32.7% vs 9.8%). Both groups struggled most on comments with a neutral Crowd opinion, however this was more significantly true for Group 2 (42.7% vs 10.7%). Both of these provide further evidence to support *Hypothesis 4* that the background of the reader of the comment has a strong impact on their ability to correctly label it.

4.4 Length

In Table 6 the correlation of security researchers (Group 1) and typical Internet users (Group 2) are presented for comments authored by a bot and a human. These results suggest that the length of a comment has some correlation with its probability of being labelled correctly, though it depends on the reader's background and the comment's author.

For Group 1, longer automatically generated comments had an increased probability of being labelled correctly ($r = 0.39$), and there was no significant correlation for comments written by a human. For Group 2, there was no significant correlation for comments by either author. Given the simplicity of the model used to generate comments, it is surprising that length had minimal impact on Group 2's probability of correctly labelling a bot comment. In comparison, Group 1 was able to apply their knowledge and capitalise on the fact that Markov chains are prone to generating text that does not make sense – particularly in long outputs – to correctly label bot-authored comments.

Table 6: Correlation between the length of a comment and its probability of being labelled correctly, grouped by author of the comment

Comment Author	Group 1	Group 2
Bot	0.39 ± 0.08	0.06 ± 0.08
Human	0.05 ± 0.11	-0.03 ± 0.02
Average	0.29 ± 0.06	0.07 ± 0.04

To an extent, this supports *Hypothesis 3* whereby the length of a comment has an impact on its ability to convince a user that it was written by a human. Surprisingly, the length of a genuine comment had no impact on the success of it being labelled correctly.

5. REFLECTIONS

Beyond the hypotheses outlined in Section 3.2, and their corresponding research questions, the goal of our work was to investigate the ease with which a bot could deceive a human into believing that an automatically generated sample of text was written by a human.

From our results we conclude that not only is it possible for a bot to deceive a human, it can be done using a simple Markov chain as a text generator. This is important as it means that our findings – for example, that security researchers and typical Internet users can be deceived more than 25% and 50% of the time respectively – are a lower bound on what is possible. Indeed, a determined actor with resources, such as a controlling government or an ambitious corporation, would be able to achieve a significantly higher success rate using more sophisticated methods and perform large scale deception.

One such example of large scale deception is the case of the online dating website Ashley Madison. The recent data leak, and the articles which followed, raised the question of whether a majority of the women on the site were actually bots [12], and if so then the company was deceiving a significant proportion of their 37 million members. A related finding from our work is that automated text on Adult content is significantly easier to deceive about than other content, even when accounting for user background. This lends support to claims that some of the 'women' on Ashley Madison could have been bots and operated successfully.

Another less publicised example is the ongoing use of social bots to deceive users into accepting their friend requests [4]. Upon accepting, these bots can harvest a user's private data, such as email addresses and phone numbers, to sell on for use in identity theft. To help combat this, and the wider issue of social bots, our findings can be used to help raise awareness of the ease with which bots can be created to deceive users. Additionally, our finding that typical Internet users are twice as likely to be deceived as security researchers highlights that there exists a knowledge gap. As such, there is some understanding which can be taught to typical Internet users that can help them to discern automatically generated content from genuine content. Furthermore, we

envisage these same factors can be used to inform the design of algorithms to detect automated online deception.

6. CONCLUSION AND FUTURE WORK

In this work, the ease with which a bot could deceive a human using automated text was investigated, motivated by the rise in use of bots for a variety of deceptive purposes. In particular, we looked at deceiving humans into believing that automatically generated text was written by a human, as well as the factors which affect how convincing the text was. We focused on the Topic and length of the text, as well as how users assess its content (Crowd opinion). For each of these, we compared the results of typical Internet users and security researchers. To accomplish this, we trained several Markov chain models on comments from Reddit to generate human-looking text that was subsequently labelled by a panel of judges.

We found that automated text is twice as likely to deceive Internet users than security researchers. Also, text that disagrees with the Crowd's opinion increases the likelihood of deception by up to 78%, while text on light-hearted Topics such as Entertainment increases the likelihood by up to 85%. Notably, we found that automated text on Adult content is the most deceptive for both typical Internet users and security researchers, increasing the likelihood of deception by at least 30% compared to other Topics on average. Together, this shows that it is feasible for a party with technical resources and knowledge to create an environment populated by bots that could successfully deceive users.

For future work, we will conduct a larger investigation into the factors which make automated text convincing. This will involve training Recurrent Neural Networks [16] to generate comments on more specific topics to be labelled by a diverse panel of judges. Additionally, we will design a realistic evaluation framework to identify when a user has been deceived by a bot which does not require the user to be explicitly told that a comment might not be genuine. Furthermore, using our gained understanding of what factors affect automated online deception, we will seek to investigate approaches and algorithms that could assist and educate users in the detection of automatically generated content.

7. REFERENCES

[1] Internet Archive. Complete Public Reddit Comments Corpus (2015). https://archive.org/details/2015_reddit_comments_corpus. Accessed 2015-09-04.

[2] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilmaz. Relevance assessment: are judges exchangeable and does it matter. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 667–674. ACM, 2008.

[3] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[4] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu. The socialbot network: when bots socialize for fame and money. In *Proceedings of the 27th Annual Computer Security Applications Conference*, pages 93–102. ACM, 2011.

[5] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Who is tweeting on twitter: human, bot, or cyborg? In *Proceedings of the 26th annual computer security applications conference*, pages 21–30. ACM, 2010.

[6] C. Clerwall. Enter the robot journalist: Users' perceptions of automated content. *Journalism Practice*, 8(5):519–531, 2014.

[7] J. P. Dickerson, V. Kagan, and V. Subrahmanian. Using sentiment to detect bots on twitter: Are humans more opinionated than bots? In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 620–627. IEEE, 2014.

[8] J. S. Donath et al. Identity and deception in the virtual community. *Communities in cyberspace*, 1996:29–59, 1999.

[9] J. Dzieza. In Case You Wondered, a Real Human Wrote This Column (2011). http://www.nytimes.com/2011/09/11/business/computer-generated-articles-are-gaining-traction.html. Accessed 2015-09-06.

[10] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini. The rise of social bots. *arXiv preprint arXiv:1407.5225*, 2014.

[11] H. H. Khondker. Role of the new media in the arab spring. *Globalizations*, 8(5):675–679, 2011.

[12] A. Newitz. Ashley Madison Code Shows More Women, and More Bots (2015). http://gizmodo.com/ashley-madison-code-shows-more-women-and-more-bots-1727613924. Accessed 2015-09-15.

[13] R. V. Noorden. Publishers withdraw more than 120 gibberish papers (2014). http://www.nature.com/news/publishers-withdraw-more-than-120-gibberish-papers-1.14763. Accessed 2015-09-06.

[14] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[15] C. Shih. *The Facebook era: Tapping online social networks to build better products, reach new audiences, and sell more stuff*. Prentice Hall, 2009.

[16] I. Sutskever, J. Martens, and G. E. Hinton. Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1017–1024, 2011.

[17] C. Wagner, S. Mitter, C. Körner, and M. Strohmaier. When social bots attack: Modeling susceptibility of users in online social networks. *Making Sense of Microposts (# MSM2012)*, 2012.

[18] R. Wald, T. M. Khoshgoftaar, A. Napolitano, and C. Sumner. Which users reply to and interact with twitter social bots? In *IEEE 25th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 135–144. IEEE, 2013.

[19] J. Xie, C. Zhang, M. Wu, and Y. Huang. Influence inflation in online social networks. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 435–442. IEEE, 2014.

[20] J. Zhang, R. Zhang, Y. Zhang, and G. Yan. On the impact of social botnets for spam distribution and digital-influence manipulation. In *IEEE Conference on Communications and Network Security (CNS)*, pages 46–54. IEEE, 2013.